Wiktionary and NLP: Improving synonymy networks

Emmanuel Navarro
IRIT, CNRS &
Université de Toulouse
navarro@irit.fr

Franck Sajous
CLLE-ERSS, CNRS &
Université de Toulouse
sajous@univ-tlse2.fr

Bruno Gaume
CLLE-ERSS & IRIT, CNRS &
Université de Toulouse
gaume@univ-tlse2.fr

Laurent Prévot
LPL, CNRS &
Université de Provence
laurent.prevot@lpl-aix.fr

Hsieh ShuKai
English Department
NTNU, Taiwan
shukai@gmail.com

Kuo Tzu-Yi
Graduate Institute of Linguistics
NTU, Taiwan
tzuyikuo@ntu.edu.tw

Pierre Magistry
TIGP, CLCLP, Academia Sinica,
GIL, NTU, Taiwan
pmagistry@gmail.com

Huang Chu-Ren
Dept. of Chinese and Bilingual Studies
Hong Kong Poly U. , Hong Kong.
churenhuang@gmail.com

Abstract

Wiktionary, a satellite of the Wikipedia initiative, can be seen as a potential resource for Natural Language Processing. It requires however to be processed before being used efficiently as an NLP resource. After describing the relevant aspects of Wiktionary for our purposes, we focus on its structural properties. Then, we describe how we extracted synonymy networks from this resource. We provide an in-depth study of these synonymy networks and compare them to those extracted from traditional resources. Finally, we describe two methods for semi-automatically improving this network by adding missing relations: (i) using a kind of semantic proximity measure; (ii) using translation relations of Wiktionary itself.

Note: The experiments of this paper are based on Wiktionary's dumps downloaded in year 2008. Differences may be observed with the current versions available online.

1 Introduction

Reliable and comprehensive lexical resources constitute a crucial prerequisite for various NLP tasks. However their building cost keeps them rare. In this context, the success of the Princeton WordNet (PWN) (Fellbaum, 1998) can be explained by the quality of the resource but also by the lack of serious competitors. Widening this observation to more languages only makes this observation more acute. In spite of various initiatives, costs make resource development extremely slow or/and result in non freely accessible resources. Collaborative resources might bring an attractive solution to this difficult situation. Among them Wiktionary seems to be the perfect resource for building computational mono-lingual and multi-lingual lexica. This paper focuses therefore on Wiktionary, how to improve it, and on its exploitation for creating resources.

In next section, we present some relevant information about Wiktionary. Section 3 presents the lexical graphs we are using and the way we build them. Then we pay some attention to evaluation (§4) before exploring some tracks of improvement suggested by Wiktionary structure itself.

2 Wiktionary

As previously said, NLP suffers from a lack of lexical resources, be it due to the low-quality or non-existence of such resources, or to copyrights-related problems. As an example, we consider French language resources. Jacquin et al. (2002) highlighted the limitations and inconsistencies from the French EuroWordnet. Later, Sagot and Fišer (2008) explained how they needed to recourse to PWN, BalkaNet (Tufis, 2000) and other resources (notably Wikipedia) to build WOLF, a free French WordNet that is promising but still a very preliminary resource. Some languages are straight-off purely under-resourced.

The *Web as Corpus* initiative arose (Kilgarriff and Grefenstette, 2003) as an attempt to design tools and methodologies to use the web for *overcoming data sparseness* (Keller and Lapata, 2002). Nevertheless, this initiative raised non-trivial technical problems described in Baroni et al. (2008). Moreover, the web is not structured enough to easily and massively extract semantic relations.

In this context, Wiktionary could appear to be a paradisiac playground for creating various lexi-

cal resources. We describe below the Wiktionary resource and we explain the restrictions and problems we are facing when trying to exploit it. This description may complete few earlier ones, for example Zesch et al. (2008a).

2.1 Collaborative editing

Wiktionary, the lexical companion to Wikipedia, is *a collaborative project to produce a free-content multilingual dictionary*.[1] As the other Wikipedia's satellite projects, the resource is not experts-led, rather filled by any kind of users. The might-be inaccuracy of the resulting resource has lengthily been discussed and we will not debate it: see Giles (2005) and Britannica (2006) for an illustration of the controversy. Nevertheless, we think that Wiktionary should be less subject (so far) than Wikipedia to voluntary misleading content (be it for ideological, commercial reasons, or alike).

2.2 Articles content

As one may expect, a Wiktionary article[2] may (not systematically) give information on a word's part of speech, etymology, definitions, examples, pronunciation, translations, synonyms/antonyms, hypernyms/hyponyms, etc.

2.2.1 Multilingual aspects

Wiktionary's multilingual organisation may be surprising and not always meet one's expectations or intuitions. Wiktionaries exist in 172 languages, but we can read on the English language main page, *"1,248,097 entries with English definitions from over 295 languages"*. Indeed, a given wiktionary describes the words in its own language but also foreign words. For example, the English article *moral* includes the word in English (adjective and noun) and Spanish (adjective and noun) but not in French. Another example, *boucher*, which does not exist in English, is an article of the English wiktionary, dedicated to the French noun *(a butcher)* and French verb *(to cork up)*.

A given wiktionary's *'in other languages'* left menu's links, point to articles in other wiktionaries describing the word in the current language. For example, the *Français* link in the *dictionary* article of the English wiktionary points to an article in the French one, describing the English word *dictionary*.

[1] http://en.wiktionary.org/

[2] What *article* refers to is more fuzzy than classical *entry* or *acceptance* means.

2.2.2 Layouts

In the following paragraph, we outline wiktionary's general structure. We only consider words in the wiktionary's own language.

An entry consists of a graphical form and a corresponding article that is divided into the following, possibly embedded, sections:

- **etymology** sections separate homonyms when relevant;
- among an etymology section, different **parts of speech** may occur;
- **definitions** and **examples** belong to a part of speech section and may be subdivided into **subsenses**;
- **translations**, **synonyms/antonyms** and **hypernyms/hyponyms** are linked to a given part of speech, with or without subsenses distinctions.

In figure 1 is depicted an article's layout example.

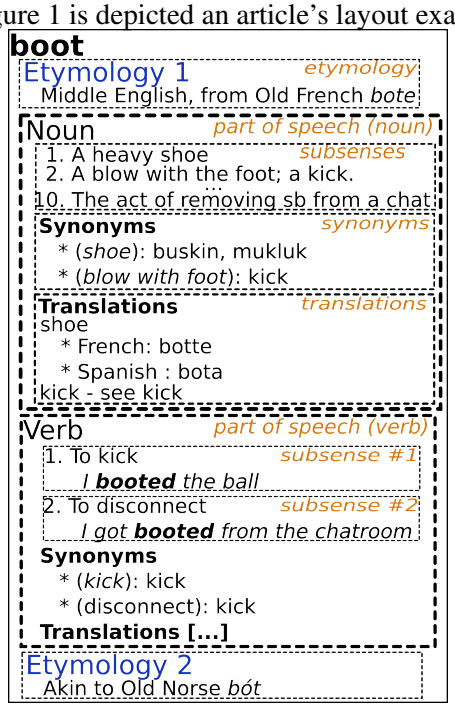

Figure 1: Layout of *boot* article (shortened)

About subsenses, they are identified with an index when first introduced but they may appear as a plain text semantic feature (without index) when used in relations (translations, synonyms, etc.). It is therefore impossible to associate the relations arguments to subsenses. Secondly, subsense index appears only in the current word (the source of the relation) and not in the target word's article it is linked to (see *orange* French N. and Adj., Jan. 10, 2008[3]).

A more serious issue appears when relations are shared by several parts of speech sections. In Ital-

[3] http://fr.wiktionary.org/w/index.php?title=orange&oldid=2981313

ian, both synonyms and translations parts are common to all words categories (see for example *cardinale* N. and Adj., Apr. 26, 2009[4]).

2.3 Technical issues

As Wikipedia and the other Wikimedia Foundation's projects, the Wiktionary's content management system relies on the MediaWiki software and on the wikitext. As stated in Wikipedia's MetaWiki article, *"no formal syntax has been defined"* for the MediaWiki and consequently it is not possible to write a 100% reliable parser.

Unlike Wikipedia, no HTML dump is available and one has to parse the Wikicode. Wikicode is difficult to handle since wiki templates require handwritten rules that need to be regularly updated. Another difficulty is the language-specific encoding of the information. Just to mention one, the target language of a translation link is identified by a 2 or 3 letters ISO-639 code for most languages. However in the Polish wiktionary the complete name of the language name (*angielski*, *francuski*, ...) is used.

2.4 Parsing and modeling

The (non-exhaustive) aforementioned list of difficulties (see §2.2.2 and §2.3) leads to the following consequences:

• Writing a parser for a given wiktionary is possible only after an in-depth observation of its source. Even an intensive work will not prevent all errors as long as (i) no syntax-checking is made when editing an article and (ii) flexibility with the "tacitly agreed" layout conventions is preserved. Better, *flexibility* is presented as a characteristic of the framework:

> *"[...] it is not a set of rigid rules. You may experiment with deviations, but other editors may find those deviations unacceptable, and revert those changes. They have just as much right to do that as you have to make them.*[5]"

Moreover, a parser has to be updated every new dump, as templates, layout conventions (and so on) may change.

• Writing parsers for different languages is not a simple adjustment, rather a complete overhaul.

• When extracting a network of semantic relations from a given wiktionary, some choices are more driven by the wiktionary inner format than scientific modelling choices. An illustration follows in §3.2. When merging information extracted from several languages, the homogenisation of the data structure often leads to the choice of the poorest one, resulting in a loss of information.

2.5 The bigger the better?

Taking advantage of colleagues mastering various languages, we studied the wiktionary of the following languages: French, English, German, Polish and Mandarin Chinese. A first remark concerns the size of the resource. The official number of declared articles in a given wiktionary includes a great number of meta-articles which are not word entries As of April 2009, the French wiktionary reaches the first rank[6], before the English one. This can be explained by the automated import of public-domain dictionaries articles (*Littré 1863* and *Dictionnaire de l'Académie Française 1932-1935*). Table 1 shows the ratio between the total number of articles and the "relevant" ones (numbers based on year 2008 snapshots).

	Total	**Meta***	**Other****	**Relevant**	
fr	**728,266**	25,244	369,948	337,074	**46%**
en	**905,963**	46,202	667,430	192,331	**21%**
de	**88,912**	7,235	49,672	32,005	**36%**
pl	**110,369**	4,975	95,241	10,153	**9%**
zh	**131,752**	8,195	112,520	1,037	**0.7%**

* templates definitions, help pages, user talks, etc.
** other languages, redirection links, etc.

Table 1: Ratio of "relevant" articles in wiktionaries

By "relevant", we mean an article about a word in the wiktionary's own language (e.g. not an article about a French word in the English Wiktionary). Among the "relevant" articles, some are empty and some do not contain any translation nor synonym link. Therefore, before deciding to use Wiktionary, it is necessary to compare the amount of extracted information contribution and the amount of work required to obtain it .

3 Study of synonymy networks

In this section, we study synonymy networks built from different resources. First, we introduce some general properties of lexical networks (§3.1). Then we explain how we build Wiktionary's synonymy network and how we analyse its properties. In §3.3, we show how we build similar graphs from traditional resources for evaluation purposes.

3.1 Structure of lexical networks

In the following sections, a graph $G = (V, E)$ is defined by a set V of n vertices and a set $E \subset V^2$ of m edges. In this paper, V is

[4] http://it.wiktionary.org/w/index.php?title=cardinale&oldid=758205

[5] http://en.wiktionary.org/wiki/WT:ELE

[6] http://meta.wikimedia.org/wiki/List_of_Wiktionaries

a set of words and E is defined by a relation $E \overset{R}{\longmapsto} E : (w_1, w_2) \in E$ if and only if $w_1 \overset{R}{\rightarrow} w_2$.

Most of lexical networks, as networks extracted from real world, are small worlds (SW) networks. Comparing structural characteristics of wiktionary-based lexical networks to some standard resource should be done according to well-known properties of SW networks (Watts and Strogatz, 1998; Barabasi et al., 2000; Newman, 2003; Gaume et al., 2008). These properties are:

• **Edge sparsity:** SW are sparse in edges $m = O(n)$ or $m = O(n\ log(n))$

• **Short paths:** in SW, the average path length (L)[7] is short. Generally there is at least one short path between any two nodes.

• **High clustering:** in SW, the clustering coefficient (C) that expresses the probability that two distinct nodes adjacent to a given third one are adjacent, is an order of magnitude higher than for Erdos-Renyi (random) graphs: $C_{SW} \gg C_{random}$; this indicates that the graph is locally dense, although it is globally sparse.

• **Heavy-tailed degree distribution:** the distribution of the vertices incidence degrees follows a power law in a SW graph. The probability $P(k)$ that a given node has k neighbours decreases as a power law, $P(k) \approx k^a$ (a being a constant characteristic of the graph). Random graphs conforms to a Poisson Law.

3.2 Wiktionary's network

Graph extraction Considering what said in §2.2.2 and §2.4, we made the following choices:[8]

• **Vertices:** a vertex is built for each entry's part of speech.

• **Parts of speech:** when modeling the links from X (X having for part of speech Pos_X) to one of its synonyms Y, we assume that $Pos_Y = Pos_X$, thus building vertex Pos_Y `.Y`.

• **Subsenses:** subsenses are flattened. First, the subsenses are not always mentioned in the synonyms section. Second, if we take into account the subsenses, they only appear in the source of the relation. For example, considering in figure 1 the relation *boot* $\xrightarrow{syn}$ *kick* (both nouns), and given the 10 subsenses for *boot* and the 5 ones for *kick*, we should build 15 vertices. And we should then add all the links between the mentioned *boot*'s subsenses and the 5 *kick*'s existing subsenses. This would lead to a high number of edges, but the graph would not be closer to the reality. The way subsenses appear in Wiktionary are unpredictable. "Subsenses" correspond sometimes to homonyms or clear-cut senses of polysemous words, but can also correspond to facets, word usage or regular polysemy. Moreover, some entries have no subsenses distinction whereas it would be worthy. More globally, the relevance of discrete word senses has been seriously questioned, see (Victorri and Fuchs, 1996) or (Kilgarriff, 1997) for very convincing discussions. Two more practical reasons led us to this choice. We want our method to be reproducible for other languages and some wiktionaries do not include subsenses. At last, some gold standard resources (eg. Dicosyn) have their subsenses flattened too and we want to compare the resources against each other.

• **Edges:** wiktionary's synonymy links are oriented but we made the graph symmetric. For example, *boot* does not appear in *kick*'s synonyms. Some words even appear as synonyms without being an entry of Wiktionary.

From the *boot* example (figure 1), we extract vertices {`N.boot`, `V.boot`}, build {`N.buskin`, `N.kick`, `V.kick`} and we add the following (symmetrized) edges: `N.boot`↔`N.buskin`, `N.boot`↔`N.kick` and `V.boot`↔`V.kick`.

Graph properties By observing the table 2, we can see that the graphs of synonyms extracted from Wiktionary are all typical small worlds. Indeed their l_{lcc} remains short, their C_{lcc} is always greater or equal than 0.2 and their distribution curves of the vertices incidence degree is very close to a power law (a least-square method gives always exponent $a_{lcc} \approx -2.35$ with a confidence r^2_{lcc} always greater than 0.89). It can also be seen that the average incidence k_{lcc} ranges from 2.32 to 3.32.[9] It means that no matter which language

[7] Average length of the shortest path between any two nodes.

[8] These choices can clearly be discussed from a linguistic point of view and judged to be biased. Nevertheless, we adopted them as a first approximation to make the modelling possible.

[9] It is noteworthy that the mean incidence of vertices is almost always the same (close to 2.8) no matter the graph size is. If we assume that all wiktionary's graphs grow in a similar way but at different speed rates (after all it is the same framework), graphs (at least their statistical properties) from different languages can be seen as snapshots of the same graph at different times. This would mean that the number of graphs edges tends to grow proportionally with the number of vertices. This fits with the dynamic properties of small worlds (Steyvers and Tenenbaum, 2005). It means that for a wiktionary system, even with many contributions, graph density is likely to remain constant and we will see that in comparison to traditional lexical resources this density is quite low.

graph	n	m	n_{lcc}	m_{lcc}	k_{lcc}	l_{lcc}	C_{lcc}	a_{lcc}	r^2_{lcc}
fr-N	18017	9650	3945	4690	2.38	10.18	0.2	-2.03	0.89
fr-A	5411	2516	1160	1499	2.58	8.86	0.23	-2.04	0.95
fr-V	3897	1792	886	1104	2.49	9.84	0.21	-1.65	0.91
en-N	22075	11545	3863	4817	2.49	9.7	0.24	-2.31	0.95
en-A	8437	4178	2486	3276	2.64	8.26	0.2	-2.35	0.95
en-V	6368	3274	2093	2665	2.55	8.33	0.2	-2.01	0.93
de-N	32824	26622	12955	18521	2.86	7.99	0.28	-2.16	0.93
de-A	5856	6591	3690	5911	3.2	6.78	0.24	-1.93	0.9
de-V	5469	7838	4574	7594	3.32	5.75	0.23	-1.92	0.9
pl-N	8941	4333	2575	3143	2.44	9.85	0.24	-2.31	0.95
pl-A	1449	731	449	523	2.33	7.79	0.21	-1.71	0.94
pl-V	1315	848	601	698	2.32	5.34	0.2	-1.61	0.92

n: number of vertices — m: number of edges
k: avg. number of neighbours per vertex — l: avg. path length between vertices
C: clustering rate — a: power law exponent with r^2 confidence
$\_{lcc}$: denotes on *largest connected component*

Table 2: Wiktionary synonymy graphs properties

or part of speech, $m = O(n)$ as for most of SW graphs (Newman, 2003; Gaume et al., 2008).

3.3 Building synonymy networks from known standards

WordNet There are many possible ways for building lexical networks from PWN. We tried several methods but only two of them are worth to be mentioned here. The graphs we built have words as vertices, not synsets or senses. A first straightforward method (method A) consists in adding an edge between two vertices only if the corresponding words appear as elements of the same synset. This method produced many disconnected graphs of various sizes. Both the computational method we planned to use and our intuitions about such graphs were pointing towards a bigger graph that would cover most of the lexical network.

We therefore decided to exploit the hypernymy relation. Traditional dictionaries indeed propose hypernyms when one look for synonyms of very specific terms, making hypernymy the closest relation to synonymy at least from a lexicographic viewpoint. However, adding all the hypernymy relations resulted in a network extremely dense in edges with some vertices having a high number of neighbours. This was due to the tree-like organisation of WordNet that gives a very special importance to higher nodes of the tree.

In the end we retained method B that consists in adding edges in following cases:

• if two words belong to the same synset;

• if a word only appears in a synset that is a leaf of the tree and contains only this word, then create edges linking to words included in the hypernym(s) synset.

Therefore when a vertice w do not get any neighbour according to method A, method B adds edges linking w to words included in the hypernym(s) synset of the synset $\{w\}$. We only added hypernyms for the leaves of the tree in order to keep our relations close to the synonymy idea. This idea has already been exploited for some WordNet-based semantic distances calculation taking into account the depth of the relation in the tree (Leacock and Chodorow, 1998).

Dicosyn graphs Dicosyn is a compilation of synonym relations extracted from seven dictionaries (Bailly, Benac, Du Chazaud, Guizot, Lafaye, Larousse and Robert):[10] there is an edge $r \rightarrow s$ if and only if r and s have the same syntactic category and at least one dictionary proposes s being a synonym in the dictionary entry r. Then, each of the three graphs (Nouns, Verbs, Adjectives) obtained is made symmetric (*dicosyn-fr-N, dicosyn-fr-V and dicosyn-fr-A*).

Properties of the graphs extracted Table 3 sums-up the structural properties of the synonyms networks built from standard resources.

We can see that all the synonymy graphs extracted from PWN or Dicosyn are SW graphs. Indeed their l_{lcc} remains short, their C_{lcc} is always greater or equal than 0.35 and their distribution curves of the vertices incidence degree is very close to a power law (a least-square method gives always exponent a_{lcc} near of -2.30 with a confidence r^2_{lcc} always greater than 0.85). It can also be observed that no matter the part of speech, the average incidence of Dicosyn-based graphs is always lower than WordNet ones.

We would like to study the evolution through time of wiktionaries, however this is outside the scope of this paper.

[10] Dicosyn has been first produced at ATILF, before being corrected at CRISCO laboratory. (`http://elsap1.unicaen.fr/dicosyn.html`)

graph	n	m	n_{lcc}	m_{lcc}	k_{lcc}	l_{lcc}	C_{lcc}	a_{lcc}	r^2_{lcc}
pwn-en-N-A	117798	104929	12617	28608	4.53	9.89	0.76	-2.62	0.89
pwn-en-N-B	117798	168704	40359	95439	4.73	7.79	0.72	-2.41	0.91
pwn-en-A-A	21479	22164	4406	11276	5.12	9.08	0.75	-2.32	0.85
pwn-en-A-B	21479	46614	15945	43925	5.51	6.23	0.78	-2.09	0.9
pwn-en-V-A	11529	23019	6534	20806	6.37	5.93	0.7	-2.34	0.87
pwn-en-V-B	11529	40919	9674	39459	8.16	4.66	0.64	-2.06	0.91
dicosyn-fr-N	29372	100759	26143	98627	7.55	5.37	0.35	-2.17	0.92
dicosyn-fr-A	9452	42403	8451	41753	9.88	4.7	0.37	-1.92	0.92
dicosyn-fr-V	9147	51423	8993	51333	11.42	4.2	0.41	-1.88	0.91

Table 3: Gold standard's synonymy graphs properties

4 Wiktionary graphs evaluation

Coverage and global SW analysis By comparing tables 2 and 3, one can observe that:

• The lexical coverage of Wiktionary-based synonyms graphs is always quantitatively lower than those of standard resources although this may change. For example, *to horn* (in PWN), absent from Wiktionary in 2008, appeared in 2009. At last, Wiktionary is more inclined to include some class of words such as *to poo* (childish) or *to prefetch, to google* (technical neologisms).

• The average number of synonyms for an entry of a Wiktionary-based resource is smaller than those of standard resources. For example, common synonyms such as *to act/to play* appear in PWN and not in Wiktionary. Nevertheless, some other appear (rightly) in Wiktionary: *to reduce/to decrease*, *to cook/to microwave*.

• The clustering rate of Wiktionary-based graphs is always smaller than those of standard resources. This is particularly the case for English. However, this specificity might be due to differences between the resources themselves (Dicosyn vs. PWN) rather than structural differences at the linguistic level.

Evaluation of synonymy In order to evaluate the quality of extracted synonymy graphs from Wiktionary, we use recall and precision measure. The objects we compare are not simple sets but graphs ($G = (V; E)$), thus we should compare separately set of vertices (V) and set of edges (E). Vertices are words and edges are synonymy links. Vertices evaluation leads to measure the resource coverage whereas edges evaluation leads to measure the quality of the synonymy links in Wiktionary resource.

First of all, the global picture (table 4) shows clearly that the lexical coverage is rather poor. A lot of words included in standard resources are not included yet in the corresponding wiktionary resources. Overall the lexical coverage is always lower than 50%. This has to be kept in mind while looking at the evaluation of relations shown in table 5. To compute the relations evaluation, each resource has been first restricted to the links between words being present in each resource.

About PWN, since every link added with method A will also be added with method B, the precision of Wiktionary-based graphs synonyms links will be always lower for "method A graphs" than for "method B graphs". Precision is rather good while recall is very low. That means that a lot of synonymy links of the standard resources are missing within Wiktionary. As for Dicosyn, the picture is similar with even better precision but very low recall.

(a) English Wiktionary vs. Wordnet

	Precision	Recall
Nouns	$14120/22075 = 0.64$	$14120/117798 = 0.12$
Adj.	$5874/8437 = 0.70$	$5874/21479 = 0.27$
Verbs	$5157/6368 = 0.81$	$5157/11529 = 0.45$

(b) French Wiktionary vs. Dicosyn

	Precision	Recall
Nouns	$10393/18017 = 0.58$	$10393/29372 = 0.35$
Adj.	$3076/5411 = 0.57$	$3076/9452 = 0.33$
Verbs	$2966/3897 = 0.76$	$2966/9147 = 0.32$

Table 4: Wiktionary coverage

5 Exploiting Wiktionary for improving Wiktionary

As seen in section 4, Wiktionary-based resources are very incomplete with regard to synonymy. We propose two tasks for adding some of these links:

Task 1: Adding synonyms to Wiktionary by taking into account its Small World characteristics for proposing new synonyms.

(a) English wiktionary vs. Wordnet

	Precision	Recall
Nouns (A)	$2503/6453 = 0.39$	$2503/11021 = 0.23$
Nouns (B)	$2763/6453 = 0.43$	$2763/18440 = 0.15$
Adj. (A)	$786/3139 = 0.25$	$786/5712 = 0.14$
Adj. (B)	$1314/3139 = 0.42$	$1314/12792 = 0.10$
Verbs (A)	$866/2667 = 0.32$	$866/10332 = 0.08$
Verbs (B)	$993/2667 = 0.37$	$993/18725 = 0.05$

(b) French wiktionary vs. Dicosyn

	Precision	Recall
Nouns	$3510/5075 = 0.69$	$3510/44501 = 0.08$
Adj.	$1300/1677 = 0.78$	$1300/17404 = 0.07$
Verbs	$899/1267 = 0.71$	$899/23968 = 0.04$

Table 5: Wiktionary synonymy links precision & recall

Task 2: Adding synonyms to Wiktionary by taking into account the translation relations.
We evaluate these two tasks against the benchmarks presented in section 3.2.

5.1 Improving synonymy in Wiktionary by exploiting its small world structure

We propose here to enrich synonymy links of Wiktionary by taking into account that lexical networks have a high clustering coefficient. Our hypothesis is that missing links in Wiktionary should be within clusters.

A high clustering coefficient means that two words which are connected to a third one are likely to be connected together. In other words neighbours of my neighbours should also be in my neighbourhood. We propose to reverse this property to the following hypothesis: "neighbour of my neighbours which are not in my neighbourhood should be a good neighbour candidate". Thus the first method we test consist simply in connecting every vertex to neighbours of its neighbours. One can repeat this operation until the expected number of edges is obtained.[11]

Secondly we used the PROX approach proposed by (Gaume et al., 2009). It is a stochastic method designed for studying "Hierarchical Small Worlds". Briefly put, for a given vertex u, one computes for all other vertices v the probability that a randomly wandering particle starting from u stands in v after a fixed number of steps. Let $P(u, v)$ be this value. We propose to connect u to the k first vertices ranked in descending order with respect of $P(u, v)$. We always choose k proportionally to the original degree of u (number of neighbours of u).

For a small number of steps (3 in our case) random wanderings tend to be trapped into local cluster structures. So a vertex v with a high $P(u, v)$ is likely to belong to the same cluster as u, which means that a link $\mathtt{u}\leftrightarrow\mathtt{v}$ might be relevant.

Figure 2 shows precision, recall and f-score evolution for French verbs graph when edges are added using "neighourhood" method (neigh), and using "Prox" method. Dashed line correspond to the value theoretically obtained by choosing edges at random. First, both methods are clearly more efficient than a random addition, which is not surprising but it seems to confirm our hypothesis that missing edges are within clusters. Adding sharply

[11] We repeat it only two times, otherwise the number of added edges is too large.

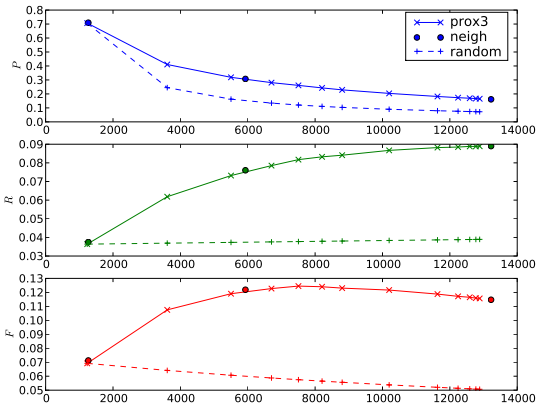

Figure 2: Precision, recall and F-score of French verbs graph enlarged using only existing synonymy links

neighbours of neighbours seems to be as good as adding edges ranked by Prox, anyway the rank provided by Prox permits to add a given number of edges. This ranking can also be useful to order potential links if one think about a user validation system. Synonyms added by Prox and absent from gold standards are not necessarily false.
For example Prox proposes a relevant link *absolve/forgive*, not included in PWN. Moreover, many false positive are still interesting to consider for improving the resource. For example, Prox adds relations such as hypernyms (*to uncover/to peel*) or inter-domain 'synonyms' (*to skin/to peel*). This is due to high clustering (see §3.1) and to the fact that clusters in synonymy networks correlates with language concepts (Gaume et al., 2008; Duvignau and Gaume, 2008; Gaume et al., 2009; Fellbaum, 1999).

Finally note that results are similar for other parts of speech and other languages.

5.2 Using Wiktionary's translation links to improve its synonymy network

Assuming that two words sharing many translations in different languages are likely to be synonymous, we propose to use Wiktionary's translation links to enhance the synonymy network of a given language.

In order to rank links to be potentially added, we use a simple Jaccard measure: let T_w be the set of a word w's translations, then for every couple of words (w, w') we have:

$$Jaccard(w, w') = \frac{|T_w \cap T_{w'}|}{|T_w \cup T_{w'}|}$$

We compute this measure for every possible pair of words and then, starting from Wiktionary's synonymy graph, we incrementally add links according to their Jaccard rank.

We notice first that most of synonymy links added by this method were not initially included in Wiktionary's synonymy network. For example, regarding English verbs, 95% of 2000 best ranked proposed links are new. Hence this method may be efficient to improve graph density. However one can wonder about the quality of the new added links, so we discuss precision in the next paragraph.

In figure 3 is depicted the evolution of precision, recall and F-score for French verbs in the enlarged graph in regard of the total number of edges. We use Dicosyn graph as a gold standard. The dashed line corresponds to theoretical scores one can expect by adding randomly chosen links.

First we notice that both precision and recall are significantly higher than we can expect from random addition. This confirms that words sharing the same translations are good synonym candidates. Added links seem to be particularly relevant at the beginning for higher Jaccard scores. From the first dot to the second one we add about 1000 edges (whereas the original graph contains 1792 edges) and the precision only decreases from 0.71 to 0.69.

The methods we proposed in this section are quite simple and there is room for improvement. First, both methods can be combined in order to improve the resource using translation links and then using clusters structure. One can also think to the corollary task that would consists in adding translation links between two languages using synonymy links of others languages.

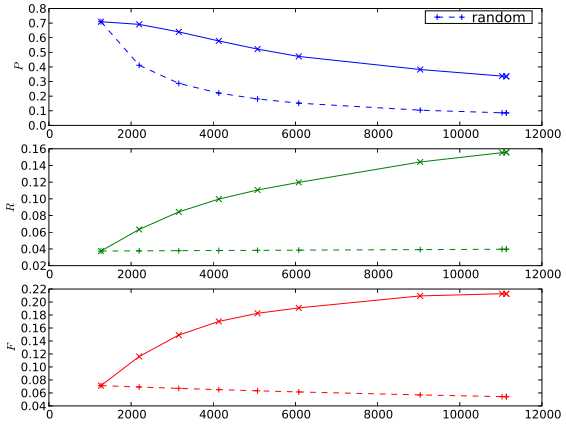

Figure 3: Precision, recall and F-score of French verbs graph enlarged using translation links

6 Conclusion and future work

This paper gave us the opportunity to share some Wiktionary experience related lexical resources building. We presented in addition two approaches for improving these resources and their evaluation. The first approach relies on the small world structure of synonymy networks. We postulated that many missing links in Wiktionary should be added among members of the same cluster. The second approach assumes that two words sharing many translations in different languages are likely to be synonymous. The comparison with traditional resources shows that our hypotheses are confirmed. We now plan to combine both approaches.

The work presented in this paper combines a NLP contribution involving data extraction and rough processing of the data and a mathematical contribution concerning graph-like resource. In our viewpoint the second aspect of our work is therefore complementary of other NLP contributions, like (Zesch et al., 2008b), involving more sophisticated NLP processing of the resource.

Support for collaborative editing Our results should be useful for setting up a more efficient framework for Wiktionary collaborative editing. We should be able to always propose a set of synonymy relations that are likely to be. For example, when a contributor creates or edits an article, he may think about adding very few links but might not bother providing an exhaustive list of synonyms. Our tool can propose a list of potential synonyms, ordered by relevancy. Each item of this list would only need to be validated (or not).

Diachronic study An interesting topic for future work is a "diachronic" study of the resource. It is possible to access Wiktionary at several stages, this can be used for studying how such resources evolve. Grounded on this kind of study, one may predict the evolution of newer wiktionaries and foresee contributors' NLP needs. We would like to set up a framework for everyone to test out new methodologies for enriching and using Wiktionary resources. Such observatory, would allow to follow not only the evolution of Wiktionary but also of Wiktionary-grounded resources, that will only improve thanks to steady collaborative development.

Invariants and variability Wiktionary as a massively mutiligual synonymy networks is an extremely promising resource for studying the (in)variability of semantic pairings such as *house/family, child/fruit, feel/know...* (Sweetser, 1991; Gaume et al., 2009). A systematic study within the semantic approximation framework presented in the paper on Wiktionary data will be carried on in the future.

References

A-L. Barabasi, R. Albert, H. Jeong, and G. Bianconi. 2000. Power-Law Distribution of the World Wide Web. *Science*, 287. (in Technical Comments).

M. Baroni, F. Chantree, A. Kilgarriff, and S. Sharoff. 2008. Cleaneval: a Competition for Cleaning Web Pages. In *Proceedings of the Conference on Language Resources and Evaluation (LREC)*, Marrakech.

Encyclopaedia Britannica. 2006. Fatally flawed: refuting the recent study on encyclopedic accuracy by the journal Nature.

K. Duvignau and B. Gaume. 2008. Between words and world: Verbal "metaphor" as semantic or pragmatic approximation? In *Proceedings of International Conference "Language, Communication and Cognition"*, Brighton.

C. Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. MIT Press.

C. Fellbaum. 1999. La représentation des verbes dans le réseau sémantique Wordnet. *Langages*, 33(136):27–40.

B. Gaume, K. Duvignau, L. Prévot, and Y. Desalle. 2008. Toward a cognitive organization for electronic dictionaries, the case for semantic proxemy. In *Coling 2008: Proceedings of the Workshop on Cognitive Aspects of the Lexicon (COGALEX 2008)*, pages 86–93, Manchester.

B. Gaume, K. Duvignau, and M. Vanhove. 2009. Semantic associations and confluences in paradigmatic networks. In M. Vanhove, editor, *From Polysemy to Semantic Change: Towards a Typology of Lexical Semantic Associations*, pages 233–264. John Benjamins Publishing.

J. Giles. 2005. Internet encyclopaedias go head to head. *Nature*, 438:900–901.

C. Jacquin, E. Desmontils, and L. Monceaux. 2002. French EuroWordNet Lexical Database Improvements. In *Proceedings of the Third International Conference on Intelligent Text Processing and Computational Linguistics (CICLING 2002)*, Mexico City.

F. Keller and M. Lapata. 2002. Using the web to overcome data sparseness. In *Proceedings of EMNLP-02*, pages 230–237.

A. Kilgarriff and G. Grefenstette. 2003. Introduction to the special issue on the web as corpus. *Computational Linguistics*, 29:333–347.

A. Kilgarriff. 1997. I don't believe in word senses. *Computers and the humanities*, 31(2):91–113.

C. Leacock and M. Chodorow. 1998. Combining local context and wordnet similarity for word sense identification. In C. Fellbaum, editor, *WordNet: An electronic lexical database*, pages 265–283. MIT Press.

M. Newman. 2003. The structure and function of complex networks.

B. Sagot and D. Fišer. 2008. Building a Free French Wordnet from Multilingual Resources. In *Proceedings of OntoLex 2008*, Marrackech.

M. Steyvers and J. B. Tenenbaum. 2005. The large-scale structure of semantic networks: Statistical analyses and a model of semantic growth. *Cognitive Science*, 29:41–78.

E. Sweetser. 1991. *From etymology to pragmatics*. Cambridge University Press.

D. Tufis. 2000. Balkanet design and development of a multilingual balkan wordnet. *Romanian Journal of Information Science and Technology*, 7(1-2).

B. Victorri and C. Fuchs. 1996. *La polysémie, construction dynamique du sens*. Hermès.

D.J. Watts and S.H. Strogatz. 1998. Collective dynamics of small-world networks. *Nature*, 393:440–442.

T. Zesch, C. Müller, and I. Gurevych. 2008a. Extracting Lexical Semantic Knowledge from Wikipedia and Wiktionary. In *Proceedings of the Conference on Language Resources and Evaluation (LREC)*, Marrakech.

T. Zesch, C. Muller, and I. Gurevych. 2008b. Using wiktionary for computing semantic relatedness. In *Proceedings of 23rd AAAI Conference on Artificial Intelligence*.